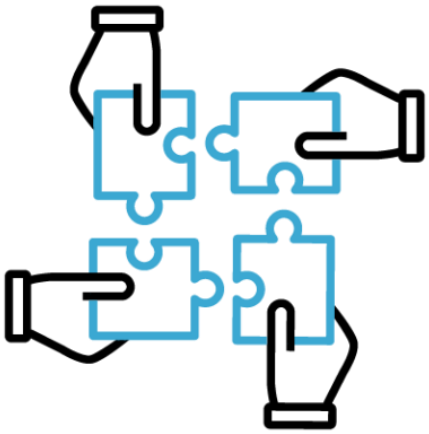
1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Anonymization of Datasets with Privacy, Utility and Risk Analysis

Assignment #1



Mestrado em Segurança Informática
Ano letivo 2022/2023

Inês Martins Marçal **Nº: 2019215917**
João Carlos Borges Silva **Nº: 2019216753**

Índice

Introdução

Este trabalho foi realizado no âmbito da cadeira de Segurança e Privacidade com o objetivo de efetuar uma análise detalhada das diversas etapas contidas no processo de anonimização de um *dataset*: caracterização do *dataset* escolhido; modelos de privacidade/anonimização selecionados e análise do risco de re-identificação que possa existir após o processo de anonimização.

Caracterização do Dataset

Descrição e objetivo

O *dataset* selecionado é composto por dados clínicos de diversos pacientes recolhidos durante exames realizados para determinar a presença ou ausência de doenças cardiovasculares. Este *dataset* contém 3 tipos de informações: objetiva, de examinação e subjetiva:

- objetiva: este tipo de categoria é inerente ao indivíduo, fundamentando-se em factos concretos e verificáveis. Exemplos deste tipo de informação poderão ser o nome e número segurança social.
- de examinação: informação obtida por meio das medições efetuadas a um determinado indivíduo. Nesta inclui-se características físicas como a pressão arterial ou o nível de açúcar no sangue.
- subjetiva: dados fornecidos pelo próprio indivíduo, podendo constituir hábitos do dia a dia ou crenças que a pessoa tenha, como por exemplo, se fuma ou não, se efetua atividade física ou a sua orientação sexual.

Pretende-se como objetivo principal anonimizar estes dados, de modo a não comprometer a privacidade dos indivíduos, preservando, simultaneamente, a sua utilidade para a criação de modelos de previsão de doenças cardiovasculares. A identificação precoce das mesmas, com base nos diversos fatores analisados, poderá permitir que os médicos consigam prevenir o desenvolvimento destas doenças.

A fim de não prejudicar este objetivo, a perda de até 4000 registos durante o processo de anonimização é considerada aceitável.

Riscos de privacidade

Visto se tratar de um *dataset* relacionado com exames médicos, é esperado que apresente riscos de privacidade para os indivíduos presentes no mesmo. A informação contida é extremamente sensível, já que a combinação de dados pessoais (nome, número de segurança social, ..) com dados médicos (doenças, estado de saúde, hábitos, …) pode permitir facilmente a identificação de um indivíduo.

Consequentemente, a identificação de uma pessoa pode levar à divulgação de ainda mais dados acerca da mesma. Por exemplo, reconhecendo um indivíduo através do nome, idade e hábitos tabágicos, torna-se possível adquirir mais informação do que aquela que se tinha inicialmente, como o número de segurança social ou que tipo de doenças a pessoa identificada tenha. A re-identificação não fica só pela divulgação de dados extra, pode ser muito mais grave que isso, pondo em causa a empregabilidade e vida social de um indivíduo, podendo inclusive levar a discriminação para com o mesmo.

A discriminação pode surgir devido, sobretudo, a informações clínicas e/ou hábitos pessoais. De seguida, apresentam-se alguns exemplos:

- Se uma pessoa estiver a contratar um seguro de saúde e a seguradora tiver acesso a informações como o nível de colesterol, a presença ou ausência de doenças cardiovasculares, entre outras, pode haver um aumento nos preços propostos, o que pode tornar esses seguros excessivamente caros.
- Da mesma forma, uma empresa poderá não contratar um determinado indivíduo caso este fume ou tenha uma vida alcoólica ativa, afetando negativamente o estado social do mesmo.

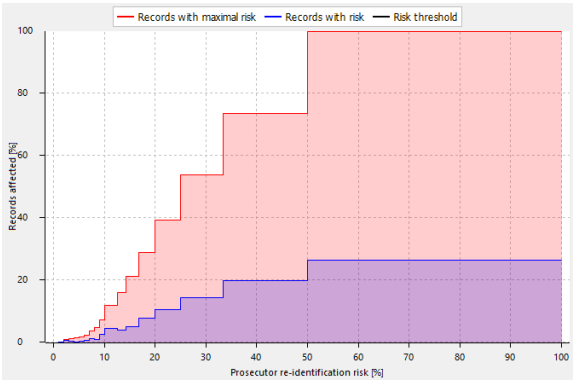
Para avaliar o risco, também é necessário considerar os diferentes modelos de ataque, já que estes serão os principais pontos para medir o sucesso do processo de anonimização no que diz respeito à privacidade. Existem 3 tipos de modelos a considerar:

- *Prosecutor*: o atacante tem apenas 1 indivíduo como alvo e sabe que o mesmo se encontra no *dataset*
- *Journalist*: a re-identificação de qualquer indivíduo beneficiará o atacante, sendo qualquer registo alvo do mesmo
- *Marketer*: neste modelo quantos mais indivíduos re-identificados melhor para o atacante. O ataque será considerado um sucesso se uma grande porção de pessoas for reidentificada.

Ao analisar os riscos de acordo com estes modelos neste *dataset*, é possível observar valores bastante elevados e perigosos para serem mantidos. Desta forma, as chances de um atacante conseguir re-identificar um indivíduo neste *dataset* são elevadas, e em alguns casos, o risco pode chegar a 100%, o que é inaceitável quando se procura manter a privacidade num *dataset*.

Prosecutor (%)			Journalist (%)			Marketer (%)
Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate
71.09857	100	47.79857	71.09857	100	47.79857	47.79857

Outro ponto importante a ser mencionado é que o risco médio é praticamente 50%, o que é um valor bastante elevado em relação ao desejado. A seguir, será apresentada a distribuição de risco:



A partir do gráfico acima, é possível ter uma visão mais clara da percentagem de registos sujeitos a serem re-identificados e perceber que ela aumenta ao longo de todo o gráfico. Com base neste mesmo gráfico, reforça-se ainda a ideia transmitida anteriormente de que é necessário realizar a anonimização deste *dataset* o mais brevemente possível.

Tratamento de colunas

O *dataset* inicial já se enquadrava no propósito deste trabalho, mas de modo a resolver algumas inconsistências, foram igualmente alteradas/acrescentadas/removidas colunas do mesmo.

Um dos primeiros problemas que o *dataset* escolhido apresentava era a falta de atributos identificadores, algo como um nome, número de telemóvel, etc. Como tal foram gerados nomes aleatoriamente assim como números de segurança social para cada utilizador, perfazendo, assim, 2 novas colunas de dados. A adição deste tipo de atributos é de extrema importância, uma vez que este trabalho visa efetuar um estudo mais aprofundado do impacto dos mesmos no processo de re-identificação.

Outra inconsistência encontrada foi o facto da coluna “idade” ser caracterizada pelo número de dias e não pelo número de anos. Isto complica o processo de anonimização, já

que torna os dados mais dispersos, e dificulta, consequentemente, a criação de hierarquias. Como tal optou-se pela sua conversão para número de anos.

Este *dataset* também apresentava alguns problemas relativamente ao intervalo de valores que as colunas “altura” e “peso” continham, alguns dos quais demasiado irrealistas (alturas com valores de 50 cm ou 250 cm, peso com valores a rondar 20 kg, pressões arteriais sistólicas/diastólicas com valores na ordem dos milhares). Como tal, estes valores foram substituídos por outros mais realistas e que se encontrassem dentro da gama esperada para um ser humano.

Por fim, foi removida a coluna "ID”, já que a mesma não apresentava qualquer utilidade para este trabalho, servindo apenas como identificador do número do registo.

Atributos

Após uma breve apresentação do *dataset*, é necessário compreender os atributos que o compõem e o significado dos mesmos, de modo a que seja possível efetuar a sua categorização de acordo com o *Privacy Preserving Data Publishing (PPDP)* e posterior agrupamento dentro das seguintes 4 categorias: identificadores, *quasi-identifying*, atributos sensíveis e atributos não sensíveis.

Este *dataset* é constituído, assim, pelos seguintes atributos: nome, número de segurança social, idade, género, altura, peso, pressão arterial sistólica, pressão arterial diastólica, nível do colesterol (normal, acima do normal ou muito acima do normal), nível de glucose (normal, acima do normal ou muito acima do normal), hábitos de tabagismo, consumo de álcool, prática de atividade física e presença ou ausência de doenças cardiovasculares.

Apresentados os atributos que vão ser utilizados durante a análise do *dataset*, é importante, como referido anteriormente, caracterizá-los de acordo com os grupos definidos pelo PPDP:

- Identificadores: identificam univocamente uma pessoa e fazem parte de um conjunto de atributos explícitos que definem um indivíduo. Neste *dataset,* os únicos atributos que apresentam estas propriedades são o nome e número de segurança social da pessoa.
- *Quasi-identifyings*: são atributos que, isoladamente, não identificam explicitamente um indivíduo, mas podem ser combinados com outros atributos ou informações adicionais para a sua re-identificação. Neste *dataset*, os atributos que pertencem a esta categoria são a idade, o género, a altura e o peso.
- Sensíveis: informação confidencial do indivíduo que jamais devem ser divulgadas devido ao caráter discriminatório que possuem, podendo de igual forma ser utilizados para re-identificação. Os atributos considerados sensíveis, neste *dataset*, são os seguintes: pressão arterial sistólica, pressão arterial diastólica, nível do colesterol, nível de glucose, hábitos de tabagismo, consumo de álcool e presença ou ausência de doenças cardiovasculares.
- Não sensíveis: são aqueles que não apresentam riscos significativos em caso de divulgação e não se enquadram em nenhum dos grupos mencionados anteriormente. Neste *dataset*, foi identificado apenas um atributo pertencente a esta categoria: a informação se o indivíduo pratica ou não atividade física.

De modo a explicitar melhor os valores possíveis que cada atributo apresenta segue-se a seguinte tabela:

Atributo	Significado de cada valor	
Nome	-	
Número social	-	
Idade		
Género	1 - mulher	2 – homem
Altura	-	
Peso	-	

Pressão arterial sistólica	-		
Pressão arterial diastólica	-		
Nível do colesterol	1 - normal	2 - acima do normal	3 - muito acima do normal
Nível da glucose	1 - normal	2 - acima do normal	3 - muito acima do normal
Hábitos de tabagismo	0 - não		1 – sim
Consumo de álcool	0 - não		1 – sim
Prática de atividade física	0 - não		1 – sim
Presença ou ausência de doenças cardiovasculares	0 - não		1 – sim

Os atributos que apresentam um "-" na coluna são autoexplicativos e, portanto, não requerem qualquer explicitação adicional sobre os respetivos valores.

Análise dos diferentes *Quasi-Identifiers*

Quando importado para o *ARX*, diferentes *quasi-identifiers (QIDs)* são sugeridos para o *dataset* em questão. A escolha do melhor *QID* depende dos valores de distinção e separação que os atributos apresentam. É preferível escolher *QIDs* com valores mais altos nestes dois campos, já que isso permite melhorar o processo de anonimização dos dados. Consideremos, por exemplo, que um *dataset* apresenta uma grande distinção de valores no atributo “idade”, mais precisamente 30 indivíduos todos com idades diferentes (entre os 20 a 50 anos). Ora, este cenário poderá favorecer a aplicação de métodos como a generalização, visto permitir a criação de três grupos de idade: [20, 30[, [30, 40[e [40, 50[.

Durante a escolha de um *QID* é necessário não só considerar os respetivos valores de distinção e separação, como também atender ao tipo de atributos escolhidos para o mesmo. Atributos identificadores, por exemplo, não devem ser selecionados como *QID*, já que podem facilmente levar à re-identificação de um indivíduo. De igual forma, não se deverá proceder à seleção de atributos sensíveis devido ao caráter discriminatório que os mesmos apresentam (tal como mencionado anteriormente), além de normalmente serem dados aos quais os indivíduos presentes no *dataset* não desejam estar associados.

Sendo assim, os atributos a serem usados como *QID* devem ser aqueles caracterizados como *quasi-identificadores,* já que são os que apresentam os melhores valores de distinção e separação em relação aos demais atributos. No entanto, é importante salientar que estes atributos devem passar por transformações, tais como a generalização ou supressão, de modo a que mais tarde seja possível criar níveis de hierarquias.

Para efetuar a escolha do melhor *QID* deverá considerar-se o número de combinações que estes atributos conseguem formar entre si, bem como os seus valores de distinção/separação. De seguida, encontram-se apresentadas, precisamente, todas as combinações possíveis entre os diferentes atributos *quasi-identifying,* bem como os respetivos valores que as mesmas apresentam:

Atributos	Distinção (%)	Separação (%)
Género	0.00286	45.4749
Idade	0.04	95.79247
Altura	0.08714	95.86133
Peso	0.16857	96.89389
Idade, Género	0.07857	97.68175
Género, Altura	0.16857	97.32633
Género, Peso	0.31143	98.248
Idade, Altura	1.90429	99.82386
Idade, Peso	2.69143	99.86939
Altura, Peso	4.17714	99.8369

Idade, Género, Altura	3.19857	99.88471
Idade, Género, Peso	4.79857	99.92546
Género, Altura, Peso	6.12714	99.89769
Idade, Altura, Peso	38.02571	99.993
Idade, Género, Altura, Peso	47.79857	99.99553

Perante os valores de distinção, torna-se evidente que *quasi-identifiers* compostos apenas por um atributo apresentam um valor reduzido. Todavia, este resultado era esperado, uma vez que o *dataset* analisado contém 70000 indivíduos e os atributos mencionados apresentam poucos valores distintos: idade tem 28, género tem 2, altura tem 61 e peso tem 118. De modo a demonstrar o processo de cálculo dos valores de distinção, será realizado um exemplo para o atributo "altura":

$$Distinção\ (\%) = \frac{número\ de\ valores\ distintos\ do\ atributo}{número\ de\ registos} \times 100 = \frac{61}{70000} \times 100 \simeq 0.08714$$

Ora, é de expectar que se o *QID* for constituído por um maior número de atributos, o número de combinações diferentes possíveis também aumenta, pelo que se irá refletir um efeito semelhante no valor da distinção. Um aspeto a ser notado é que os *QIDs* com valores mais baixos (atendendo ao número de elementos: 1, 2, 3) serão sempre aqueles que apresentem o "género", pois o mesmo demonstra ter uma distinção baixíssima comparativamente com os restantes atributos.

Quando se fala em separação todos os *QIDs*, exceto o género, apresentam um valor alto, entre os 95% e 100%. O valor baixo que o género apresenta deve-se, em parte, à sua baixa distinção, já que o mesmo apresenta apenas 2 valores (1 ou 2), o que leva a que, em caso de igualdade, seja menos provável separar 2 registos utilizando este atributo. De modo a demonstrar a aplicação da fórmula do cálculo da separação, será efetuado um exemplo para o atributo "género":

$$Separação\ (\%) = \frac{número\ de\ tuplos\ onde\ não\ há\ repetição\ de\ valores}{número\ de\ tuplos\ possíveis\ de\ formar} \times 100$$

$$= \frac{número\ de\ tuplos\ possíveis\ de\ formar - número\ de\ tuplos\ onde\ há\ repetição\ de\ valores}{número\ de\ tuplos\ possíveis\ de\ formar} \times 100$$

$$= \frac{70000_{C_2} - (45530_{C_2} + 24470_{C_2})}{70000_{C_2}} \times 100 = \left(1 - \frac{1335845900}{2449965000}\right) \times 100 \simeq 45.479$$

Sendo que os valores de separação estão praticamente todos dentro da mesma gama, o *QID* para este *dataset* deve ser escolhido com base na distinção. Desta forma, os QIDs com valores mais satisfatórios serão "idade, altura, peso" ou "idade, gênero, altura, peso", apesar de apresentarem uma diferença de 9% entre si, serão ambos aceitáveis como *quasi-identifiers*. É ainda de esperar que estes valores calculados pelo *ARX* diminuam após o processo de anonimização.

Análise da distribuição do dataset e definição dos *coding models*

Seguidamente, é analisada a distribuição do *dataset* em questão de modo a poderem ser definidos os *coding models* mais adequados para os diferentes atributos. Além disto, será ainda importante atender ao objetivo inicialmente estabelecido e à caracterização efetuada para os diferentes atributos.

É de relembrar que o objetivo definido para este *dataset* é auxiliar o desenvolvimento de modelos que permitam prever se um determinado paciente está em risco de ter doenças cardiovasculares, portanto, pode haver atributos que não apresentem qualquer utilidade para este fim. Um exemplo destes atributos são os *identifying*, já que para o objetivo estabelecido a informação relativa ao nome e número de segurança social não apresentam qualquer

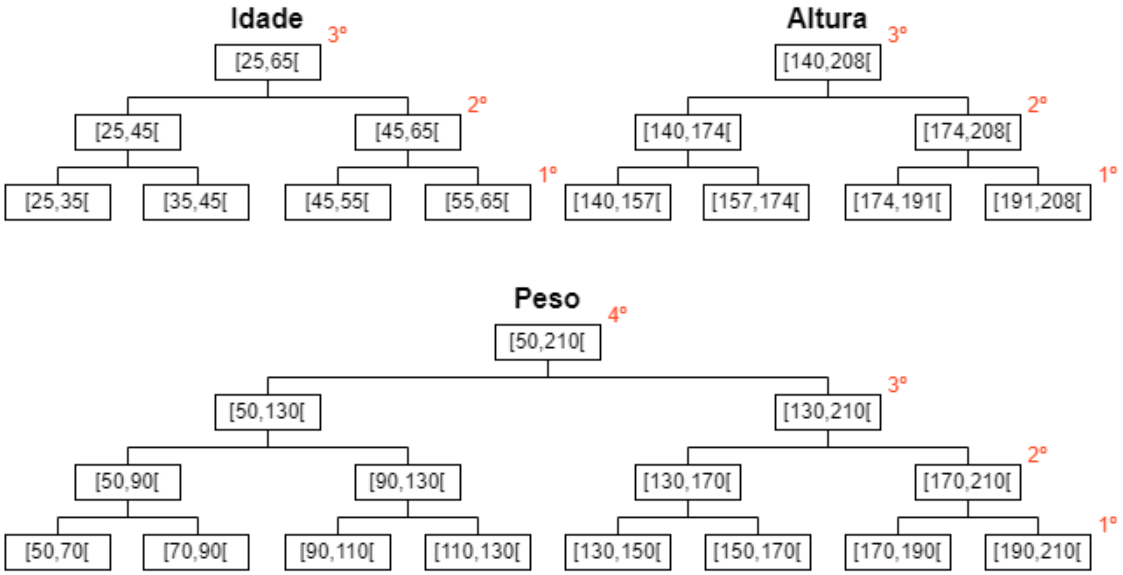
relevância na previsão deste tipo de doenças. Deste modo, procedeu-se à aplicação do *coding model* supressão nas colunas que continham este tipo de atributo.

Informação classificada como sensível não deverá ser modificada, visto que a mesma é extremamente fundamental para a obtenção do objetivo definido para este *dataset*. Por esta mesma razão não foi aplicado qualquer *coding model* a este tipo de atributos. Seguindo esta ordem de ideias, a informação fornecida por atributos como "prática de atividade física" e "género" deverá permanecer inalterada.

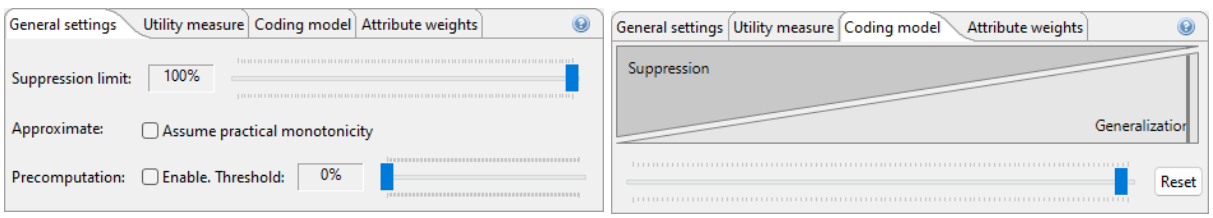
Restando os atributos "idade", "altura" e "peso", o tipo de *coding model* escolhido para os mesmos foi a generalização por hierarquia de intervalos. Esta escolha deve-se à larga gama de valores que cada um destes atributos apresenta. De modo a compreender melhor que tipo de intervalos deveriam ser definidos foi utilizado o programa *Excel* para determinar os mínimos e máximos destes atributos. Os valores obtidos foram os seguintes:

Atributo	Valor mínimo	Valor máximo
Idade	29	64
Altura	140	207
Peso	60	200

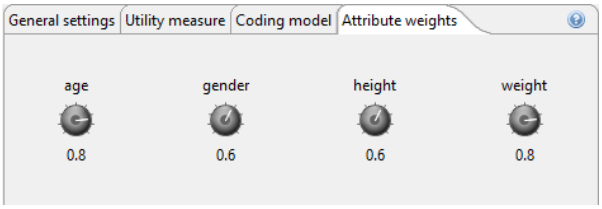
Com base nestes valores, pode-se concluir que para os atributos "idade" e "altura", a criação de 4 intervalos (3 níveis de hierarquia) constituiu a melhor opção. Quanto ao "peso" foram criados 8 intervalos (4 níveis de hierarquia), visto este apresentar uma maior gama de valores entre o valor máximo e mínimo. Procedeu-se à organização dos intervalos da seguinte forma:



Quanto à definição do *coding model* foram efetuados alguns ajustes no que toca aos seus parâmetros. Na aba *General settings*, procedeu-se ao aumento do *suppresion limit* para 100%, de modo a que fosse possível utilizar todos os níveis de hierarquia definidos anteriormente. Já na aba *Coding model*, o cursor foi ajustado de modo a estar o mais próximo possível do valor máximo de generalização, o que permitiu a criação de um maior número de combinações dos níveis de hierarquia.



Na aba *Attribute weights* foram estabelecidos diferentes pesos para os diferentes atributos, considerando que cada um destes possui um nível de importância distinto. Enquanto que a idade e o peso apresentam uma maior relevância para o objetivo definido, o género e altura têm um nível de importância semelhante, mas mais baixo A distribuição de pesos estabelecida foi a seguinte:



Finalmente, foi avaliado que para a criação dos modelos de previsão de doenças cardiovasculares seria aceitável a utilização dos seguintes níveis de hierarquia (onde ✅ significa aceitável e ❌ não aceitável):

Atributo	Nível 0	Nível 1	Nível 2	Nível 3	Nível 4
Idade	❌	✅	✅	❌	—
Altura	❌	✅	✅	❌	—
Peso	❌	✅	✅	❌	❌

A escolha destas restrições está diretamente associada ao nível de detalhe que se pretende para o desenvolvimento do objetivo inicialmente definido.

NOTA: células que apresentem — significa que este nível não existe para tal atributo

Aplicação de modelos de privacidade

Modelos escolhidos

Ao longo deste trabalho, foram realizadas várias experiências com diferentes modelos de privacidade com vista a minimizar o risco de ataques dos diversos modelos (*Prosecutor, Journalist, Marketer*) e reduzir a perda de informação para níveis praticamente insignificantes. Como tal, foram testados não só modelos que proporcionassem privacidade aos *QIDs* (*K-Anonymity* e/ou *(ε, δ)-Differential Privacy*), como também outros que permitissem melhorar e complementar os mesmos (*ℓ-Diversity, t-Closeness* e/ou *β-Likeness*).

Modelos base para todas as experiências

Inicialmente, a ideia seria começar por aplicar o modelo *K-Anonymity* e, posteriormente, avaliar o seu desempenho quando combinado com outras técnicas complementares, a fim de avaliar o quanto estes melhoravam os parâmetros anteriormente referidos (menor risco e informação perdida). Infelizmente, não é possível realizar esta experiência no *ARX*, visto ser necessário definir previamente um modelo de privacidade para os atributos sensíveis. Assim, todos os testes efetuados envolvem a aplicação do modelo *K-Anonymity* (aplicado a todos os atributos *quasi-identifiers*) em conjunto com o modelo *ℓ-Diversity* (aplicado a todos os atributos sensíveis). A razão para esta escolha está relacionada com a necessidade de proteger os indivíduos deste *dataset* contra eventuais re-identificações por parte do modelo de ataque *Prosecutor*. A adoção do modelo *ℓ-Diversity* juntamente com o *K-Anonymity* decorre do fato de o primeiro poder melhorar o desempenho do segundo. Isto verifica-se, precisamente, porque o *K-Anonymity*, na ausência de diversificação dos atributos sensíveis em uma classe de equivalência, pode não funcionar da maneira mais adequada.

Considerando estes como os modelos “base” das experiências efetuadas, combinados, posteriormente, com os restantes, será necessário definir um valor tanto para o K como para

o ℓ, de modo a evitar uma grande supressão nos dados, mas assegurando, em simultâneo, um nível de privacidade aceitável. Ora, sabe-se que a quantidade de supressão necessária para garantir a privacidade é inversamente proporcional à preservação dos dados e é com base nesta relação que se vai tentar encontrar os melhores parâmetros para os modelos *K-Anonymity* e *ℓ-Diversity*.

Escolha do ℓ para o ℓ-Diversity

Para a definição do valor ℓ, é necessário analisar a quantidade de valores distintos que cada atributo sensível possui:

Atributo	Número de valores distintos
Frequência cardíaca sistólica	118
Frequência cardíaca diastólica	72
Nível de colesterol	3
Nível da glucose	3
Hábitos de tabagismo	2
Consumo de álcool	2
Presença ou ausência de doenças cardiovasculares	2

Pode-se observar que este parâmetro tende a ser maior para os atributos "frequência cardíaca sistólica" e "frequência cardíaca diastólica", uma vez que os mesmos apresentam maior número de valores distintos relativamente a outros atributos, que possuem apenas 2 ou 3 valores distintos. Desta forma, para os restantes atributos foi aplicada a técnica *ℓ-Diversity* com um parâmetro igual ao respetivo número de valores distintos, visando melhorar o desempenho dos modelos de privacidade utilizados. De seguida, foram efetuados experimentos de modo a encontrar os melhores valores para o parâmetro ℓ a aplicar nos atributos “frequência cardíaca sistólica” e “frequência cardíaca diastólica”. É importante encontrar um valor adequado para este parâmetro, que não seja muito grande a ponto de comprometer a qualidade dos dados, nem muito pequeno a ponto de não garantir a privacidade necessária e, portanto, para estes 2 atributos começou-se por adotar os valores 20 e 15, respetivamente. A partir deste ponto, foram utilizados valores progressivamente menores para avaliar o melhor ℓ. Seguem-se os resultados dos experimentos efetuados:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1,0,1,1	0	0.42918	0.04759	0	0.42918	0.04759	0.04759	4859
1,0,1,2	0	0.29155	0.03276	0	0.29155	0.03276	0.03276	2848
1,0,2,1	0	0.33003	0.0349	0	0.33003	0.0349	0.0349	4098
1,0,2,2	0	0.2681	0.02505	0	0.2681	0.02505	0.02505	2193
2,0,1,1	0	0.33333	0.0332	0	0.33333	0.0332	0.0332	3740
2,0,1,2	0	0.27248	0.02196	0	0.27248	0.02196	0.02196	1680
2,0,2,1	0	0.32787	0.02535	0	0.32787	0.02535	0.02535	2932
2,0,2,2	0	0.26738	0.01738	0	0.26738	0.01738	0.01738	949
ℓ em fcs = 20, ℓ em fcd = 15 e ℓ dos restantes = número de valores distintos								

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1,0,1,1	0	0.68027	0.05691	0	0.68027	0.05691	0.05691	3224
1,0,1,2	0	0.42918	0.03662	0	0.42918	0.03662	0.03662	1737

1,0,2,1	0	0.63291	0.043	0	0.63291	0.043	0.043	2561
1,0,2,2	0	0.42735	0.02887	0	0.42735	0.02887	0.02887	718
2,0,1,1	0	0.57803	0.03967	0	0.57803	0.03967	0.03967	1941
2,0,1,2	0	0.51546	0.02599	0	0.51546	0.02599	0.02599	754
2,0,2,1	0	0.57143	0.0306	0	0.57143	0.0306	0.0306	1362
2,0,2,2	0	0.83333	0.02155	0	0.83333	0.02155	0.02155	399

ℓ em fcs = 15, ℓ em fcd = 15 e ℓ dos restantes = número de valores distintos

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1,0,1,1	0	0.69444	0.05938	0	0.69444	0.05938	0.05938	2641
1,0,1,2	0	0.98039	0.03803	0	0.98039	0.03803	0.03803	1635
1,0,2,1	0	0.68027	0.04557	0	0.68027	0.04557	0.04557	1970
1,0,2,2	0	0.97087	0.03027	0	0.97087	0.03027	0.03027	615
2,0,1,1	0	0.5988	0.04104	0	0.5988	0.04104	0.04104	1774
2,0,1,2	0	0.9434	0.0274	0	0.9434	0.0274	0.0274	648
2,0,2,1	0	1.08696	0.03338	0	1.08696	0.03338	0.03338	1101
2,0,2,2	0	0.83333	0.02155	0	0.83333	0.02155	0.02155	399

ℓ em fcs = 15, ℓ em fcd = 10 e ℓ dos restantes = número de valores distintos

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1,0,1,1	0	1.63934	0.06869	0	1.63934	0.06869	0.06869	1577
1,0,1,2	0	1.2987	0.04056	0	1.2987	0.04056	0.04056	972
1,0,2,1	0	2.22222	0.05367	0	2.22222	0.05367	0.05367	1064
1,0,2,2	0	1.72414	0.03168	0	1.72414	0.03168	0.03168	557
2,0,1,1	0	1.63934	0.04912	0	1.63934	0.04912	0.04912	777
2,0,1,2	0	1.2987	0.02881	0	1.2987	0.02881	0.02881	571
2,0,2,1	0	1.38889	0.03743	0	1.38889	0.03743	0.03743	541
2,0,2,2	0	0.83333	0.02155	0	0.83333	0.02155	0.02155	399

ℓ em fcs = 10, ℓ em fcd = 10 e ℓ dos restantes = número de valores distintos

A relação previamente estabelecida torna-se mais clara se se atender aos experimentos efetuados. Como foi referido, à medida que o ℓ vai aumentando, o risco irá diminuir, mas em contrapartida o número de registos removidos aumenta cada vez mais. Sendo que se dispõe de um grande número de registos, a remoção de alguns deles não afetará significativamente o objetivo estabelecido. No entanto, é importante minimizar a quantidade de dados removidos. Outro ponto a não ser menosprezado é a taxa de risco (%) que cada modelo de ataque apresenta, uma vez que é crucial manter o máximo nível de privacidade possível.

Aplicando ℓ=10 para os dois atributos, consegue-se perceber que a informação perdida é muito reduzida comparada com os restantes atributos. Acontece que a taxa de risco para este experimento pode ultrapassar o 1 ou 2%, o que fez com que esta configuração de parâmetros fosse eliminada das opções. Ao analisar a experiência em que são aplicados ℓ=20 e ℓ=15, é possível observar um comportamento oposto. O risco apresentado é bastante baixo, no entanto, o número de registos perdidos é maior, o que, tal como na experiência anteriormente analisada, levou a que esta configuração de valores fosse descartada como opção.

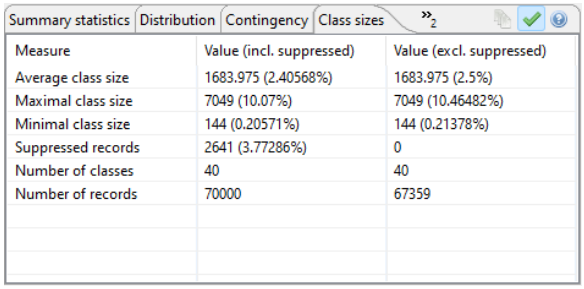
Restam apenas as experiências com as seguintes configurações: [ℓ=15 e ℓ=15] e [ℓ=15 e ℓ=10]. Apesar da 2ª opção apresentar um risco ligeiramente superior (só algumas combinações de hierarquias), o número de registos perdidos é menor. Tendo em conta que

ainda se vai aplicar mais modelos de privacidade e o risco irá diminuir, fui decidida que a melhor opção seria no final de contas a configuração ℓ em fcs = 15, ℓ em fcd = 10. Posto isto, as experiências daqui em diante terão como base a aplicação de *ℓ-Diversity* com a seguinte configuração:

	fcs	fcd	Nível de colestrol	Nível da glucose.	Hábitos de Tabagismo	Presença ou ausência de doenças cardiovasculares	Consumo de álcool
Valor de ℓ	15	10	3	3	2	2	2

Escolha do K para o K-Anonymity

Tendo escolhido o valor do ℓ, é necessário definir o valor do K para o modelo *K-Anonymity*. Perante um conjunto de valores distintos tão reduzido num *dataset* tão extenso, é de se esperar que escolher um valor pequeno para o K não ajude praticamente em nada no processo de anonimização. Os seguintes resultados demonstram, respetivamente, a *utility* (em relação a *class sizes*) apresentada sem *K-Anonymity*, quando os níveis de hierarquia são 1,0,1,1 para o *QID*:



Como se pode observar, utilizando unicamente *ℓ-Diversity* como anteriormente mencionado, o tamanho mínimo de uma classe de equivalência já é 144. Isto dá-nos indícios de que, usando um K com valor menor ou igual 144, a percentagem de risco de qualquer modelo de ataque permanecerá a mesma. A razão para este comportamento é bem simples, o *K-Anonymity* procura criar classes de equivalência com tamanho igual ou superior ao valor de K. Sendo o número de classes igual a este, a aplicação deste modelo não melhorará o nível de risco.

Com base nestes dados, foram testados K's cada vez maiores de modo a avaliar qual a melhor solução para o objetivo em questão. É de relembrar que este modelo terá o mesmo comportamento que o *ℓ-Diversity*, isto é, quanto maior o parâmetro do modelo, menor será o risco, mas maior será a perda de registos. Considerando que a combinação ideal de hierarquias é 1,0,1,1 (por deixar os valores mais próximos do valor original), foram realizados os seguintes testes utilizando esta combinação:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
K	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[145,148[	0	0.68027	0.05802	0	0.68027	0.05802	0.05802	2785
[148,157[	0	0.64103	0.05666	0	0.64103	0.05666	0.05666	2932
[157,212[	0	0.47393	0.0553	0	0.47393	0.0553	0.0553	3088
[212,229[	0	0.4386	0.05397	0	0.4386	0.05397	0.05397	3299
[229,234[	0	0.42918	0.05265	0	0.42918	0.05265	0.05265	3527
[234,254[	0	0.39526	0.05133	0	0.39526	0.05133	0.05133	3760
[254,302[	0	0.33113	0.05001	0	0.33113	0.05001	0.05001	4013

Sabendo que os próximos modelos que forem aplicados vão aumentar ainda mais o número de registos eliminados e tendo como definido eliminação no máximo de 4000 registos, escolheu-se um K que esteja no intervalo de [157,212[. Desta forma, há margem para a perda de ainda 1000 registos e a percentagem de risco será igualmente baixa. O K escolhido será então 200, por exemplo.

K-Anonymity + ℓ-Diversity + t-Closeness

Definido os parâmetros do *K-Anonymity* e *ℓ-Diversity* irá tentar-se combinar os mesmos com outros modelos de privacidade, de modo a conseguir obter um risco próximo de 0%, mas considerando sempre o menor número de registos removidos possível. O primeiro modelo de privacidade experimentado foi o *t-Closeness* e a sua escolha surge da mesma forma que o *ℓ-Diversity*. Este modelo é, assim, utilizado de modo a resolver alguns problemas que o *ℓ-Diversity* apresenta: o mesmo não considera a semântica dos valores sensíveis, nem a distribuição global dos mesmos. O *t-Closeness* procura, assim, manter a distribuição de valores sensíveis dentro das classes de equivalência o "mais próximo" da tabela original. Este "mais próximo" refere-se, portanto, a um parâmetro t.

À semelhança do que ocorreu com as outras distribuições, também será necessário gerenciar um parâmetro. Este parâmetro é um limiar para a 'distância' entre a distribuição dos atributos sensíveis nas classes de equivalência do novo *dataset* anonimizado (Q) e a distribuição original destes atributos (P) no *dataset* não anonimizado. Por outras palavras: *Distância (P, Q) <= t*.

Inicialmente, tentou-se estabelecer limites para os valores de t. Para isso, foram calculados, novamente utilizando o Excel, os limites inferiores e superiores dos valores que os atributos sensíveis podem assumir:

Atributo	Valor Mínimo	Valor Máximo
Frequência cardíaca sistólica	77	240
Frequência cardíaca diastólica	60	182
Nível de colesterol	1	3
Nível do glucose	1	3
Hábitos de tabagismo	0	1
Consumo de álcool	0	1
Presença ou ausência de doenças cardiovasculares	0	1

A partir destes mesmos valores calculou-se a distância percentual do valor mínimo ao valor máximo:

Atributo	Distância (%)	Fórmula
Frequência cardíaca sistólica	67.9	$d = 1 - \frac{valor\ mínimo}{valor\ máximo}$
Frequência cardíaca diastólica	67.0	
Nível de colesterol	66.7	
Nível do glucose	66.7	
Hábitos de tabagismo	50*	* de modo uniformizar os valores percentuais e evitar percentagens de 0% considera-se o min=1 e max=2
Consumo de álcool	50*	
Presença ou ausência de doenças cardiovasculares	50*	

Estes valores permitem, assim, estabelecer um limite superior para a escolha de um valor para t, possibilitando a aplicação do *t-Closeness* em cada atributo. De modo a avaliar a *performance,* será utilizada a mesma técnica empregada na definição dos parâmetros dos outros modelos, variando gradualmente o t. Ao contrário do que foi feito com os outros modelos, o parâmetro t irá ser progressivamente reduzido de modo a efetuar os testes, uma vez que valores maiores não trarão benefícios adicionais. Isto ocorre porque a distância percentual é um limite superior, e valores maiores não melhoram o nível de privacidade.

Posto isto, foram efetuados inicialmente alguns testes onde o t fosse igual para todos os atributos sensíveis, seguindo novamente os níveis de hierarquia 1,0,1,1:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
t	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
0.4	0	0.47393	0.0553	0	0.47393	0.0553	0.0553	3088
0.3	0	0.47393	0.05046	0	0.47393	0.05046	0.05046	4599
0.2	0	0.47393	0.04361	0	0.47393	0.04361	0.04361	12668
0.1	0	0.04156	0.02211	0	0.04156	0.02211	0.02211	60956

Como se pode observar, o risco praticamente não diminuiu, exceto a percentagem de sucesso de cada ataque. As mudanças significativas ocorrem apenas a partir de um valor de t=0.1, mas, neste caso, há uma supressão de 85% dos dados, o que é completamente inaceitável.

Deste modo, foi testado outra abordagem, desta vez decrementando-se o parâmetro t em 0.1, começando com o mesmo igual ao limiar anteriormente calculado para cada atributo. Os resultados são os seguintes:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
t	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
d - 0.1	0	0.47393	0.0553	0	0.47393	0.0553	0.0553	3088
d - 0.2	0	0.47393	0.0553	0	0.47393	0.0553	0.0553	3088
d - 0.3	0	0.47393	0.04891	0	0.47393	0.04891	0.04891	10709
d - 0.4	0	0.24938	0.03525	0	0.24938	0.03525	0.03525	52979

Novamente, os resultados ficam aquém do pretendido, pois, assim como nas primeiras experiências, a redução do valor de t não contribui em nada para a diminuição do risco de re-identificação. É ainda de notar que a supressão é cada vez maior, conduzindo a uma redução na utilidade do *dataset*.

Conclui-se, portanto, que não faz sentido aplicar a combinação deste modelo de privacidade (*t-Closeness*) juntamente com os modelos base (*K-Anonymity* e *ℓ-Diversity*) no contexto em que este *dataset* se encontra. Os testes efetuados comprovaram que tanto a supressão quanto a quantidade de privacidade não atingiram níveis satisfatórios para o objetivo proposto.

K-Anonymity + ℓ-Diversity + (ε,δ)-Differential Privacy

O próximo modelo que se tentou combinar com os modelos base foi o *(ε,δ)-Differential Privacy*. Este foca-se na ideia de adicionar ruído ao *dataset* de modo a preservar algumas características do *dataset* original, mas aumentando a privacidade do mesmo. Como o ruído adicionado é aleatório, o risco de ser identificado pelos diferentes modelos de ataque pode variar de execução para execução.

O presente modelo é gerido por dois parâmetros ε e δ, que, por sua vez, apresentam papéis um pouco diferentes. O primeiro controla a quantidade de ruído adicionado, tornando mais difícil a identificação de indivíduos, enquanto que o segundo controla a probabilidade máxima aceitável de que a privacidade de um indivíduo seja violada. É importante referir que

quanto menores forem estes valores, mais privacidade o *dataset* ganhará, porém, em contrapartida, apresentará mais ruído, ou seja, tornará mais difícil alcançar o objetivo inicialmente proposto.

Portanto, como sempre, haverá um *tradeoff* entre a quantidade de informação perdida e o nível de privacidade alcançável. Os primeiros testes foram efetuados fixando o ε no valor *default* do *ARX* (2.0) e variando o valor do δ. É ainda necessário realçar que o próprio *ARX* já sugere quais os níveis de hierarquia a aplicar a cada elemento do *QID*, não sendo possível alterar os mesmos. Com base nesta ideia apresenta-se a seguinte tabela de resultados:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
δ, ε=2	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1×10^{-4}	0	0.3937	0.05979	0	0.33113	0.05015	0.05015	3314
1×10^{-5}	0	0.39216	0.05975	0	0.33113	0.05001	0.04996	3355
1×10^{-6}	0	0.39062	0.05988	0	0.33113	0.05001	0.04998	3704
1×10^{-7}	0	0.39062	0.05858	0	0.33113	0.04882	0.04877	3346
1×10^{-8}	0	0.39062	0.05988	0	0.33113	0.05001	0.04994	3348

Os resultados são praticamente iguais para qualquer δ, pelo que, neste caso, optou-se por utilizar o de menor valor. Um aspeto a ter em atenção é que os resultados apresentados representam os riscos mínimos encontrados e podem variar em cada execução devido à adição aleatória de ruído, conforme mencionado anteriormente.

Para tentar encontrar um valor ideal para o ε utilizou-se a mesma estratégia, ou seja, fixou-se o δ (no valor 1x10^-8) e variou-se o ε, de modo a eventualmente encontrar um valor satisfatório.

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
$\delta=10^{-8}$, ε	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[0.3,1[	0	0.39027	0.05996	0	0.33113	0.04961	0.04994	3248
0.2	0	0.38462	0.1212	0	0.06423	0.01903	0.02012	1710
0.1	0	0.35461	0.11553	0	0.06423	0.01903	0.01903	1953

Novamente, é possível perceber que os resultados são praticamente idênticos para parâmetros ligeiramente diferentes, como é o caso de ε's que estejam no intervalo [0.3,1[. Ao tentar utilizar um valor de ε igual a 0.2, observa-se uma ligeira diminuição na taxa de risco. No entanto, é importante notar que a combinação de hierarquias muda de 1,0,1,1 para 1,0,1,2. Utilizando um valor como 0.1 para este parâmetro, o risco é ainda mais baixo, mas desta vez a combinação de hierarquias passa a ser 1,0,1,3. Assim, o valor ótimo para ε é 0.2. Embora não tenha sido possível obter valores de risco mais baixos utilizando a hierarquia 1,0,1,1, o conjunto de resultados obtidos através deste valor de ε é muito próximo em termos de utilidade em relação ao *dataset* original, exigindo apenas um nível adicional de hierarquia para o atributo "peso".

Conclui-se, portanto, que, ao contrário do modelo *t-Closeness*, o modelo *(ε,δ)-Differential Privacy* apresenta resultados positivos para valores de ε=0.2 e δ=10^-8. Isto permitiu não só diminuir o risco apresentado pelos modelos base, mas também reduzir a quantidade de supressão necessária no *dataset*.

K-Anonymity + ℓ-Diversity + β-Likeness

Por último, tentou-se explorar a utilização do modelo *β-Likeness* juntamente com os modelos base. A definição dada para este modelo é a seguinte: "Dado um *dataset* DB com o atributo sensível *SA*, considere-se V = {v1, . . . , vm} como domínio de *SA*, e P = (p1, . . . , pm) a distribuição geral de SA no *Dataset*. Uma classe de equivalência G com distribuição de SA Q = (q1, . . . , qm) satisfaz *β-likeness*, se e só se $max\{D(pi, qi)|pi \in P, pi < qi\} \leq \beta$, onde *β*

> *0* e *D(pi, qi)* é a distância entre *pi* e *qi* que é calculada por $\frac{q_i - p_i}{p_i}$ " . É de frisar que um *dataset* só cumpre isto, se todas as classes de equivalência seguirem precisamente esta definição. Está-se, assim, perante mais um método que visa limitar a distância máxima entre a distribuição original e a distribuição anonimizada.

Com base na definição acima é expectável que quanto menor o β, maior deverá ser a privacidade, uma vez que o modelo estará menos preocupado em assegurar a semelhança entre o *dataset* original e o anonimizado. Por outro lado, quanto menor este parâmetro maior será a supressão de registos e como sempre é necessário fazer um balanceamento entre estas duas medidas. Desta forma, foram testados diversos parâmetros para os diferentes atributos sensíveis separadamente, ou seja, *K-Anonymity* + *ℓ-Diversity* + (*β-Likeness* em apenas um atributo sensível). Os seguintes resultados representam os melhores *tradeoffs* obtidos entre perda de informação e minimização do risco, sendo o β apresentado o melhor valor para esta relação e combinação de níveis de hierarquia 1,0,1,1:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
β (atributo)	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
27 (fcs)	0	0.4386	0.04886	0	0.4386	0.04886	0.04886	8600
8 (fcd)	0	0.22779	0.05417	0	0.22779	0.05417	0. 05417	34924
0.5 (Nível de colest.)	0	0.39526	0.05019	0	0.39526	0.05019	0.05019	18195
0.5 (Nível de gluc.)	0	0.39526	0.04357	0	0.39526	0.04357	0.04357	12626
1.8 (Habitos de taba.)	0	0.4386	0.04863	0	0.4386	0.04863	0.04863	6254
1.8 (Consumo de álcool)	0	0.4386	0.05397	0	0.4386	0.05397	0.05397	3299
0.1 (Presença ou ausência de doenças cardio.)	0	0.39526	0.03868	0	0.39526	0.03868	0.03868	49318

Como pode ser observado, o único atributo onde realmente existe um bom *tradeoff* é o do "consumo de álcool", a aplicação de *β-Likeness* em qualquer outro atributo torna-se pejorativa. Dado isto, é de esperar que ao se tentar aplicar este modelo a mais do que um atributo, o resultado da privacidade até pode vir a aumentar, mas o nível de supressão será tão elevado que o *dataset* ficará inutilizável.

Pode-se concluir que a melhor opção para aplicar o modelo *β-Likeness* neste *dataset* é usar somente o valor β=1.8 para o atributo "consumo de álcool".

K-Anonymity + ℓ-Diversity + β-Likeness + (ε,δ)-Differential Privacy

Sendo *β-Likeness* (com β=1.8 apenas no atributo "consumo de álcool") e *(ε,δ)-Differential Privacy* os únicos modelos que melhoraram a performance do *K-Anonymity* + *ℓ-Diversity*, resolveu-se testar a aplicação conjunta destes 2 modelos com os modelos base. É importante relembrar que a aplicação do *(ε,δ)-Differential Privacy* implica a adição de ruído aleatório. A seguir, são apresentadas algumas execuções realizadas combinando estes quatro modelos:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Número de registos apagados
Top melhores experiências	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1º (1,0,1,2)	0	0.34602	0.11568	0	0.05285	0.01895	0.01895	1908

2º (1,0,1,2)	0	0.35714	0.12186	0	0.06423	0.02023	0.02023	1675
3º (1,0,1,2)	0	0.37037	0.12265	0	0.06423	0.0201	0.01994	1715
4º (1,0,1,2)	0	0.37175	0.12252	0	0.06423	0.0201	0.02004	1606
5º (1,0,1,2)	0	0.37313	0.12221	0	0.06423	0.02041	0.02041	1620

Perante os melhores resultados, percebe-se que em geral o risco diminuiu um pouco mais, enquanto a supressão aumentou muito pouco, sem prejudicar a *utility* do *dataset*. Ainda assim, à semelhança do que foi observado durante a aplicação do *(ε,δ)-Differential Privacy*, a combinação de níveis de hierarquia que resultou na redução de risco foi 1,0,1,2, não representando, como já referido, uma ameaça para o objetivo definido.

Análise geral do nível de utilidade, risco e privacidade

Utilidade

O objetivo inicial era desenvolver um modelo de previsão de doenças cardiovasculares. Portanto, é necessário avaliar o quão próximo se está deste objetivo, mesmo após todo o processo de anonimização. O *dataset* desempenha um papel fundamental neste processo, pois é com base no mesmo que serão criados os modelos preditivos. Para tal, deverá ainda proceder-se à divisão do mesmo em duas partes: treino e teste, podendo-se definir cada uma destas da seguinte forma:

- treino: são o tipo de registos que devem ser utilizados durante a construção e afinação do modelo preditivo. Estas informações serão a base para o modelo classificar se uma pessoa tem ou não doenças cardiovasculares
- teste: tipo de registos utilizados para avaliar a exatidão do modelo. Durante o teste, o modelo é aplicado a estes registos e produz uma previsão sobre a presença ou ausência de doenças cardiovasculares. Estas, por sua vez, são então comparadas com os resultados reais para avaliar a capacidade do modelo de efetuar previsões precisas.

A medida de utilidade, neste caso, deverá ser a quantidade de informação perdida, onde quanto menos registos forem perdidos pelo processo de anonimização melhor será a utilidade. De seguida será apresentada uma visão geral da quantidade de informação perdida em cada processo de anonimização como consequência da aplicação dos modelos de privacidade:

Modelo aplicado	Número de registos máximos perdidos	Combinação de hierarquias
K-Anonymity + ℓ-Diversity	3088	1,0,1,1
K-Anonymity + ℓ-Diversity + t-Closeness	3088	1,0,1,1
K-Anonymity + ℓ-Diversity + (ε,δ)-Differential Privacy	1710	1,0,1,2
K-Anonymity + ℓ-Diversity + β-Likeness	3299	1,0,1,1
K-Anonymity + ℓ-Diversity + β-Likeness + (ε,δ)-Differential Privacy	1908	1,0,1,2

Quanto ao número de registos perdidos, as combinações de modelos de privacidade que utilizam o modelo *(ε,δ)-Differential Privacy* são as que demonstram melhores resultados, sendo também as que menos riscos apresentam (como vai ser analisado no próximo capítulo). Contudo, a combinação de hierarquias não é a mais ideal, mas ainda assim apresenta uma grande utilidade, já que adicionar mais um nível de hierarquia no atributo "peso" terá pouco impacto no desenvolvimento do modelo preditivo.

Foi, assim, desenvolvido um modelo preditivo utilizando a técnica de regressão logística, em Python, de modo a avaliar a eficácia dos *datasets* anonimizados em alcançar o objetivo definido. Primeiramente, deverá submeter-se o *dataset* exportado a algumas alterações:

exclusão das linhas do arquivo *.csv* (registos) que foram suprimidas e remoção das colunas correspondentes aos atributos "nome" e "número de segurança social", suprimidas de igual forma.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# read dataset
df = pd.read_csv('TREINO.csv', sep=';')

# drop non relevant columns
df.drop('name', axis=1, inplace=True)
df.drop('social_number', axis=1, inplace=True)

# drop suprassed lines
m = 3088
df = df.drop(df.index[-m:], axis=0)
```

Visto que as funções utilizadas em *Python* para este fim apenas trabalham com valores numéricos, as colunas que, ao serem anonimizadas, foram transformadas em intervalos ("idade", "altura" e "peso") precisam ser convertidas para tal. Desta forma, para atingir o objetivo proposto, estas colunas são transformadas em outras. Por exemplo, a coluna "idade" pode ter os valores [25,35[, [35,45[, [45,55[e [55,65[, pelo que será representada através das novas colunas com os nomes "[25,35[", "[35,45[", "[45,55[" e "[55,65[", respetivamente. Estas, por sua vez, terão o valor 1 ou 0, dependendo do valor inicial da idade. As antigas colunas como “idade”, “altura” e “peso” foram, assim, substituídas.

```
# conver intervals to numeric variables
interval_dummies_age = pd.get_dummies(df["age"])
interval_dummies_height = pd.get_dummies(df["height"])
interval_dummies_weight = pd.get_dummies(df["weight"])

df.drop('age', axis=1, inplace=True)
df.drop('height', axis=1, inplace=True)
df.drop('weight', axis=1, inplace=True)
# reagroup the dummie variables
X = pd.concat([df.iloc[:,:-1], interval_dummies_age], axis=1)
X = pd.concat([df.iloc[:,:-1], interval_dummies_height], axis=1)
X = pd.concat([df.iloc[:,:-1], interval_dummies_weight], axis=1)
y = df["cardio"]
```

De seguida, o restante *dataset* é dividido nos conjuntos de treino e teste, com 20% dos dados escolhidos para teste e 80% para treino:

```
# create the train and test sets of data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Finalmente, aplica-se a técnica de regressão logística e calcula-se algumas métricas de avaliação do modelo.

```
# apply logistic regression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

y_pred = logreg.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
confusion = confusion_matrix(y_test, y_pred)
classification = classification_report(y_test, y_pred)
print("Accuracy:", accuracy)
print("Confusion Matrix:\n", confusion)
print("Classification Report:\n", classification)
```

A seguinte tabela permite visualizar a precisão, *recall*, *f1-score* e exatidão obtidas por meio dos diferentes *datasets* gerados pelas diversas combinações de modelos de privacidade:

Modelo aplicado	Precisão	Recall	F1-Score	Exatidão
K-Anonymity + ℓ-Diversity	0.76	0.65	0.70	0.72
K-Anonymity + ℓ-Diversity + t-Closeness	0.76	0.65	0.70	0.72
K-Anonymity + ℓ-Diversity + (ε,δ)-Differential Privacy	0.74	0.65	0.69	0.72
K-Anonymity + ℓ-Diversity + β-Likeness	0.77	0.65	0.70	0.72
K-Anonymity + ℓ-Diversity + β-Likeness + (ε,δ)-Differential Privacy	0.77	0.65	0.71	0.73

Com base nos dados acima, consegue-se perceber que o modelo preditivo construído a partir do *dataset* previamente anonimizado apresenta uma exatidão de cerca de 72%. Este valor é considerado bastante satisfatório para o objetivo em questão visto demonstrar que a utilidade do *dataset* permanece praticamente intacta, mesmo após todo o processo de anonimização.

Risco e privacidade

Além de avaliar a utilidade, é importante considerar o risco e a privacidade dos *datasets* anonimizados. Para isso, o *ARX* apresenta um conjunto de estatísticas que permitem facilitar esta análise: distinção e separação dos diferentes *QIDs*, risco segundo os diferentes modelos de ataque e distribuição do risco.

Começando por analisar a percentagem de distinção e separação dos *QIDs*, é importante destacar que, após o processo de anonimização, espera-se que a mesma tenha diminuído relativamente ao *dataset* original. Ora, isto é vantajoso porque, por exemplo, quanto menos distintos forem os *QIDs*, mais difícil será para um atacante re-identificar um determinado indivíduo num *dataset* anonimizado. Analisemos a distinção e separação dos diferentes *QIDs* segundo as diferentes combinações de modelos:

QID	1º	2º	3º	4º	5º
P	0.006/58.91	0.003/23.38	0.02/10.96	0.003/23.38	0.02/11.09
G	0.003/45.10	0.003/45.47	0.02/42.86	0.003/45.47	0.02/42.58
A	0.004/42.14	0.001/0	0.03/33.57	0.001/0	0.03/33.19

I	0.004/61.04	0.004/61.35	0.03/59.12	0.004/61.35	0.03/59.02
G, P	0.010/77.20	0.005/57.55	0.03/50.21	0.005/57.55	0.03/50.05
A, P	0.015/76.11	0.003/23.38	0.04/44.04	0.003/23.38	0.04/43.81
G, A	0.007/66.34	0.003/45.47	0.04/62.10	0.003/45.47	0.04/61.57
I, P	0.016/84.03	0.009/70.52	0.05/63.98	0.009/70.52	0.05/63.94
I, G	0.009/78.52	0.009/78.84	0.06/76.47	0.009/78.84	0.06/76.39
I, A	0.013/77.59	0.004/61.35	0.07/74.56	0.004/61.35	0.07/74.34
G, A, P	0.021/86.11	0.005/57.55	0.05/67.73	0.005/57.55	0.05/67.34
I, G, P	0.029/91.11	0.017/83.60	0.08/79.75	0.017/83.60	0.08/79.70
I, A, P	0.040/90.77	0.009/70.52	0.09/78.17	0.009/70.52	0.09/78.01
I, G, A	0.019/86.92	0.009/78.84	0.10/84.80	0.009/78.84	0.10/84.58
I, G, A, P	0.055/94.62	0.017/83.60	0.12/87.20	0.017/83.60	0.12/87.05
P - Peso, G – Género, A – Altura, I - Idade					

Em comparação com os valores iniciais, observa-se que a distinção e separação diminuíram, o que indica uma redução no risco de re-identificação. Esta diminuição de valores era expectável, já que, após a anonimização, o número valores distintos diminui, o que leva a uma menor distinção e separação, seguindo a fórmula aplicada aos mesmos.

De seguida, considerando o risco dos diferentes modelos de ataques, é esperado novamente que o mesmo diminua, já que com a aplicação de diferentes modelos de privacidade, maior poderá vir a ser o tamanho das classes de equivalência. Atentando à fórmula do risco máximo, segundo o modelo *Prosecutor*, fica ainda mais evidente a relação entre o tamanho das classes e o risco:

$$P_{Prosecutor} = \frac{1}{F_j}\ , F_j \text{ é o tamanho da classe de equivalência } j$$

Os seguintes valores já foram comentados durante a aplicação das diferentes combinações de modelos, mas, de modo a ser possível efetuar uma melhor comparação entre estes, juntou-se o risco mais baixo obtido por cada modelo de ataque na seguinte tabela:

	Prosecutor (%)			Journalist (%)			Marketer (%)	Combinação de hierarquias aplicada
Combinação de modelos	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
1º	0	0.47393	0.0553	0	0.47393	0.0553	0.0553	1,0,1,1
2º	0	0.47393	0.0553	0	0.47393	0.0553	0.0553	1,0,1,1
3º	0	0.38462	0.1212	0	0.06423	0.01903	0.02012	1,0,1,2
4º	0	0.4386	0.05397	0	0.4386	0.05397	0.05397	1,0,1,1
5º	0	0.34602	0.11568	0	0.05285	0.01895	0.01895	1,0,1,1

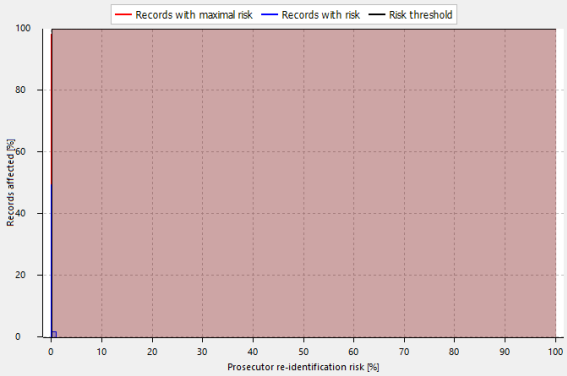
1º - K-Anonymity + ℓ-Diversity
2º - K-Anonymity + ℓ-Diversity + t-Closeness
3º - K-Anonymity + ℓ-Diversity + (ε,δ)-Differential Privacy
4º - K-Anonymity + ℓ-Diversity + β-Likeness
5º - K-Anonymity + ℓ-Diversity + β-Likeness + (ε,δ)-Differential Privacy

Os parâmetros para cada experiência são aqueles que foram considerados os melhores durante a análise dos diferentes modelos de privacidade

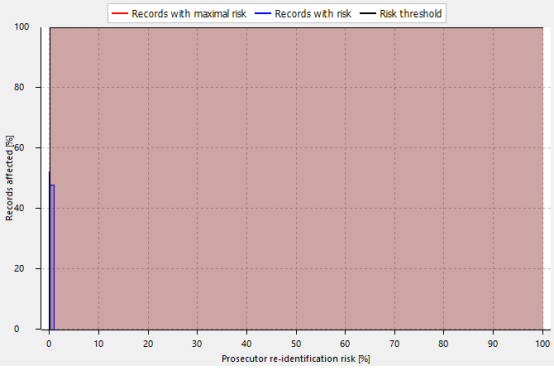
Os valores dos diferentes riscos mostram-se bastante satisfatórios comparado com os valores iniciais, perfazendo, assim, uma diminuição de 98% do risco em grande parte das combinações dos modelos. Outro aspeto interessante de realçar é que mesmo com um risco máximo tão baixo, a percentagem de registos que apresenta o mesmo é pouca. Por fim, se considerarmos a média de risco, o resultado é novamente muito bom em comparação com a média original, o que é outro aspeto bastante positivo deste processo de anonimização.

Combinação de modelos	Lowest prosecutor risk (%)	Registos afetados pelo Lowest prosecutor risk (%)	Average prosecutor risk (%)	Registos afetados pelo Highest prosecutor risk (%)
1º	0.01419	10.53	0.0553	0.32
2º	0.01419	10.53	0.0553	0.32
3º	0.04897	20.96	0.12322	2.79
4º	0.01419	10.57	0.05397	0.34
5º	0.04778	21.57	0.12371	2.85

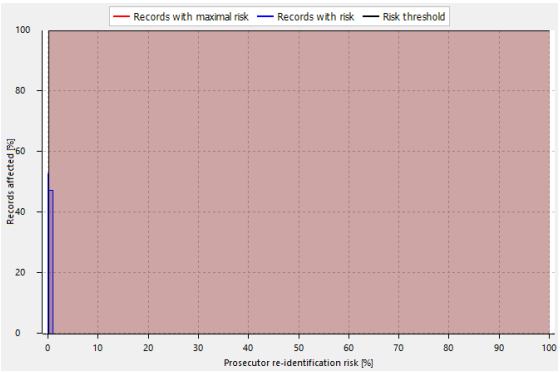
Por último, deve-se atender à distribuição do risco ao longo dos registos do *dataset*, retratada nos seguintes gráficos:



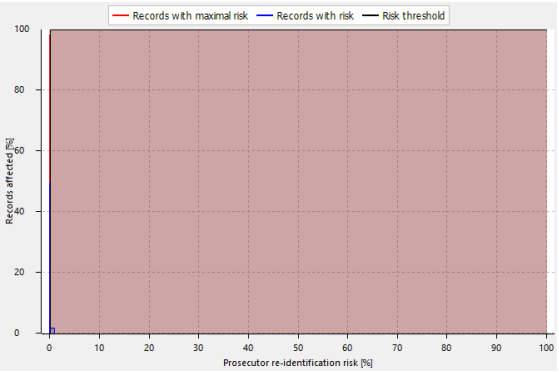
1ª combinação



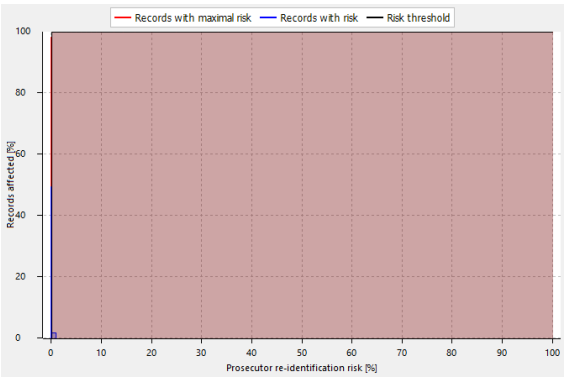
2ª combinação



3ª combinação



4ª combinação



5ª combinação

O risco permanece, assim, constante na maioria dos registos, mas visto que o risco máximo apresenta um valor bastante reduzido, isto não contribuirá para um maior perigo de re-identificação.

À semelhança do que foi observado na utilidade, os resultados na quantidade de riscos e privacidade apresentam novamente valores bastante satisfatórios, em comparação com o *dataset* original. Ainda assim, desta vez é possível distinguir um pouco melhor qual combinação de modelos apresenta melhor impacto no *dataset*, podendo-se concluir que a combinação *K-Anonymity + ℓ-Diversity + β-Likeness + (ε,δ)-Differential Privacy*, apresenta os melhores valores do que toca ao risco e privacidade.

Conclusão

O presente trabalho permitiu, portanto, fornecer uma visão mais clara dos passos que devem ser tomados no processo de anonimização de um *dataset.* Foi importante desde o início definir e compreender os requisitos de privacidade que o *dataset* resultante deveria seguir e interligar estes requisitos com o objetivo para o qual o *dataset* está destinado, bem como com o tipo de dados com os quais se está lidando.

Começou-se por efetuar uma caracterização completa dos dados iniciais, atribuindo uma classe a cada atributo (identificador, *quasi-identifying*, sensível, não sensível). Além disso, foi importante analisar a distribuição dos dados para uma melhor definição dos *coding models*, como hierarquias de generalização, supressão, pesos, entre outros.

Após este passo, foram estudados que modelos de privacidade fariam mais sentido serem aplicados no *dataset* em questão. Dentro de cada modelo, foram analisados quais os melhores parâmetros a atribuir, de modo a que houvesse sempre um melhor *tradeoff* entre a utilidade e o nível de risco/privacidade exigidos. Um ponto importante utilizado para fundamentar as escolhas efetuadas foi revisitar a análise da distribuição do *dataset* e, por vezes, refazê-la.

Finalmente, foi analisado se os objetivos anteriormente definidos tanto para a utilidade como para a privacidade foram cumpridos. Concluiu-se que os resultados foram, em ambos os casos, satisfatórios, demonstrando, assim, o bom estudo realizado sobre o *dataset* utilizado. Como combinação de modelos de privacidade destacou-se a combinação de *K-Anonymity + ℓ-Diversity + β-Likeness + (ε,δ)-Differential Privacy* com os devidos parâmetros previamente descritos.

Referências

- Slides da disciplina de Segurança e Privacidade
- https://arx.deidentifier.org/anonymization-tool/configuration/#a222
- https://arx.deidentifier.org/anonymization-tool/exploration/
- https://arx.deidentifier.org/overview/privacy-criteria/
- https://petsymposium.org/popets/2018/popets-2018-0004.pdf
- https://cs.au.dk/~karras/pvldb5(11)_p1388.pdf
- https://www.mdsaude.com/hipertensao/pressao-arterial-normal/
- https://www.analyticssteps.com/blogs/what-differential-privacy-and-how-does-it-work
- https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset